Diagnosis and prognosis of osteoarthritis by texture analysis using sparse linear models

Marques, Joselene; Clemmensen, Line Katrine Harder; Dam, Erik

Publication date:
2012

Document version
Peer reviewed version

Diagnosis and prognosis of osteoarthritis by texture analysis using sparse linear models

Joselene Marques[1,2], Line Clemmensen[3], and Erik Dam[2]

[1]e-Science, University of Copenhagen, Denmark
[2]Biomediq, Copenhagen, Denmark
[3]Department of Informatics and Mathematical Modelling, Technical University of Denmark

Abstract. *We present a texture analysis methodology that combines uncommitted machine-learning techniques and sparse feature transformation methods in a fully automatic framework. We compare the performances of a partial least squares (PLS) forward feature selection strategy to a hard threshold sparse PLS algorithm and a sparse linear discriminant model. The texture analysis framework was applied to diagnosis of knee osteoarthritis (OA) and prognosis of cartilage loss. For this investigation, a generic texture feature bank was extracted from magnetic resonance images of tibial knee bone. The features were used as input to the sparse algorithms, which defined the best features to retain in the model. To cope with the limited number of samples, the data was evaluated using 10 fold cross validation (CV). The diagnosis evaluation using sparse PLS reached a generalization area-under-the-ROC curve (AUC) of 0.93 and the prognosis had AUC of 0.70, both superior to established cartilage based markers known to relate to OA diagnosis and prognosis.*

Keywords: sparse PLS, sparse LDA, sparsity, feature selection, texture analysis, OA, bone structure

1 Introduction

Osteoarthritis (OA) is a complex disease that affects multiple components of the joint. Up to 80% of the population over 65 years old suffer from OA symptoms, leading to an impaired quality of life [1]. Following similar investigations [2], texture analysis on MRI images of patients with OA may capture early tissue changes and provide the means for obtaining information that might not be assessed visually.

Some texture analysis approaches combine different features at various scales in a generic feature bank to allow, for instance, a broad representation of the image [3, 4]. The drawback of a feature bank is a potentially high-dimensional representation of data, which usually have a large number of correlated features. Besides providing nearly the same information to predict the classes, these features can imply model convergence problems and overfitting. In such situations, sparse methods can reduce the non relevant features by adding an appropriate

penalty term to the objective function. The induced sparsity have the potential of yielding a simplified and more interpretable model of the scientific problem been investigated.

Different forms of the penalty terms have been proposed in the literature [5]. The Ridge regression minimizes a penalized objective function by adding an L2 penalty term, which shrinks the values towards zero. The Lasso adds an L1 penalty to the objective function. The resulting method can be seen as a variable selection strategy, since some of the estimated variables are forced to zero, depending on the size of this penalty. The Elastic Net combines L1 and L2 penalty, selecting some variables such as Lasso and shrinking some variables according to Ridge.

In this study, we investigated dimensionality reduction (DR) by sparsity methods. We used a texture analysis framework applied to diagnosis of knee OA and prognosis of cartilage loss to compare the performance of the DR strategies using a sparse partial least squares (PLS) method and a sparse linear discriminant analysis (LDA) method.

2 Background

LDA is a well known classification strategy for low-dimensional problems. However, in high dimensions, LDA can result in poor performance due to high variance. In this context, data analysis techniques like PCA and PLS have the common principle of finding a linear transformed space that potentially allows space reduction without loosing important information.

For n samples and p features, PCA maximizes the variance of the data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ based on the decomposition of $\mathbf{X}'\mathbf{X}$. In this approach, no importance is given to how each feature may be related to the classes in $\mathbf{Y} \in \mathbb{R}^{n \times 1}$. In another way, the PLS-based approach suggests the use of supervised dimensionality reduction by considering both $\mathbf{X}$ and $\mathbf{Y}$. The PLS regression model computes the coefficient matrix by successive 1-D linear regression, so it does not require matrix inversion, allowing the estimation of the relationship between features and classes when the number of features exceeds the number of samples.

2.1 Dimensionality reduction using linear discriminant analysis

LDA can be derived using different approaches. The Fisher LDA estimates a low-dimensional discriminative space defined by linear transformations that maximizes the ratio of between-class scatter to within-class scatter. In an alternative approach, the optimal scoring implementation recasts the classification problem as a regression problem. The categorical variables are turned into quantitative variables by defining $\mathbf{Y}$ as an $n \times j$ matrix of dummy variables for the j classes and n observations. By linear regression, the algorithm assigns scores to the classes, where the coefficient matrix reflect the optimal scores. See more details in [6].

Though LDA often performs quite well in low-dimensional data, it is known to fail when the number of features is larger than the number of observations [7]. In this case, LDA cannot be applied directly, without some regularization.

To deal with the high dimensions, the sparse LDA presented in [7] applies an Elastic Net penalty to the coefficient vectors in the optimal scoring interpretation of LDA. Besides performing classification and feature selection simultaneously, the imposed sparseness criterion of this approach allows to set the exact number of non-zero loadings desired in each discriminative direction.

Our present investigation evaluates the above sparse strategy as an alternative implementation of the DR step in the texture analysis framework.

2.2 Dimensionality reduction using partial least squares

Succinctly, a PLS regression model decomposes $\mathbf{X}$ and $\mathbf{Y}$ to produce the bilinear representation of the data presented in equations 1 and 2. The $\mathbf{X}$-scores, $\mathbf{T} \in \mathbb{R}^{n \times h}$, contain the h transformed features in the orthogonal space, while the matrix $\mathbf{U} \in \mathbb{R}^{n \times h}$ has the transformed classes. $\mathbf{P} \in \mathbb{R}^{p \times h}$ and $\mathbf{Q} \in \mathbb{R}^{1 \times h}$ are the loading matrices. The matrices $\mathbf{E} \in \mathbb{R}^{n \times p}$ and $\mathbf{F} \in \mathbb{R}^{n \times 1}$ contain the residuals.

The latent factors (equation 3) and regression coefficients (equation 4) are computed based on a weight matrix $\mathbf{W} \in \mathbb{R}^{p \times h}$ that expresses the correlation of each $\mathbf{X}$-column with the $\mathbf{Y}$ variable. Thereby, entries in $\mathbf{W}$ with values close to zero express less important features.

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \qquad (1) \qquad\qquad \mathbf{T} = \mathbf{X}\mathbf{W} \qquad (3)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F} \qquad (2) \qquad\qquad \mathbf{B} = \mathbf{W}\mathbf{Q}' \qquad (4)$$

To estimate the features more related to the classification classes, the work in [8] introduced a DR strategy based on PLS. The presented method used the PLS output to rank the features and implemented a learning step that iteratively selected the most important features.

However, Hyonho et al. [9] showed that a large number of noise features forces the PLS loadings to divert from the direction that relates $\mathbf{X}$ and $\mathbf{Y}$, which can cause inconsistency. Considering that it can attenuate estimates of the regression parameters, the authors proposed a PLS formulation with imposed sparsity on the direction vectors. The proposed sparse method is equivalent to the Elastic Net approach, which selects some variables and shrinks some values towards zero.

Since the PLS sparse approach tends to avoid inconsistency on the direction vectors, in the present work, we investigated whether it can efficiently identify the relevant texture features for diagnosis of OA and prognosis of cartilage loss. Building on Hyonho et al.'s proposed formulation we implemented a sparse PLS algorithm and included it as a feature selection step to the texture analysis framework.

3 Application of the framework for OA diagnosis and prognosis

In this section, we detail the texture analysis framework and its application to diagnosis and prognosis of OA. The overall framework included the following ordered steps: segmentation of the region-of-interest (ROI), features computation,

dimensionality reduction, classification and evaluation. Section 3.1 introduces a brief description of the data collection and Section 3.3 explains the dimensionality reduction method including the detailed implementation of sparse PLS.

3.1 Data Collection

The data set consisted of MRI of both left and right knees from 159 test subjects in a community-based, non-treatment study (in accordance with the Helsinki Declaration II and European Guidelines for Good Clinical Practice and approved by the local ethical committee). After exclusion of scans due to acquisition artefacts, 313 knee scans were included in the diagnosis dataset. Due to subjects that dropped out in the follow-up MRI acquisition, the prognosis dataset had only 268 knee scans. The scan size was 104 X 170 X 170, after automatically removing boundaries with no information. The healthy and diseased subjects were determined by radiologists and the levels of cartilage loss were assessed by a segmentation process [10].

3.2 Data Set Generation

ROI definition: A voxel classification algorithm [10] segmented the tibial knee bone. From the segmented binary-mask image, we applied a morphological erosion of approximately 0.5 mm to remove the cortical bone (the outer layer of the bone). The remaining was the trabecular bone, which was defined as the ROI. The Figure 1 shows examples of a ROI *(a)* and knee joints healthy *(b)* and diseased *(c)*.

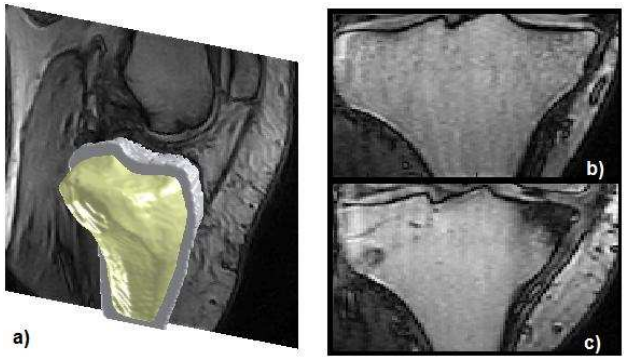

Fig. 1. a) Automatically segmented tibia bone in light gray and the region-of-interest (the trabecular bone) in gold. On the right, the figure shows an example of a healthy *(b)* and a diseased knee *(c)*.

Texture features: Following an uncommitted approach, 178 generic texture features were extracted from the images at three different scales . A feature set that has been demonstrated to provide good results for many patterns is the N-jet [11, 10, 4]. The N-jet applies Gaussian derivative filters equivalent to the partial derivatives of a local Taylor series approximation up to order N. The partial derivatives calculated at a given image point and a given scale were used as a basis for representing the different visual textures of the image. We included the 3-jet, based on the Gaussian derivative kernels including up to the third order. Furthermore, gradient vector and magnitude and non-linear combinations of the Gaussian derivative features were included.

Feature scores: The features were extracted for each voxel. In order to summarize them across a ROI, three scores were calculated: the mean, the standard deviation, and the Shannon entropy. When extracting the features at three scales and calculating the three feature scores for each ROI, the total number of features was 534.

3.3 The dimensionality reduction method

As pre-processing step, the feature set was normalized to zero mean and a standard deviation of one. Next, the sparse PLS algorithm defined the selected features and the number of PLS latent factors used in the final model.

The sparse PLS algorithm: The original implementation of the SIMPLS algorithm [12] (see the first column of Table 1) uses conjugate gradient (CG) [13] to compute the coefficients. At each iteration, one column of the matrix $\mathbf{W}$ is computed based on the correlation of each $\mathbf{X}$-column with the $\mathbf{Y}$ variable. The sparse PLS algorithm (see the second column of Table 1) applied hard threshold by imposing zero values to all elements of the $\mathbf{W}$-column with absolute value less than the specified threshold. The candidate thresholds were defined to be logarithmically spaced between the minimum and the maximum value of the first latent factor. Using cross validation, each candidate threshold was sent to the sparse PLS algorithm and evaluated. This intermediate evaluation considered different numbers of PLS latent factors. For the best evaluation, the algorithm identified the selected features by the non-zero value in the $\mathbf{W}$-matrix. Note that the threshold and number of PLS latent factors were optimized by cross validation, while the selected features were determined by the sparsity algorithm. The final model included the selected feature set and number of PLS latent factors.

The sparse LDA evaluation: As an alternative implementation of the framework, this investigation replaced the sparse PLS by the sparse LDA, in order to compare both approaches of sparsity. During the training phase, a cross validation strategy optimized the lambda and number of selected features. The best combination of theses two parameters was used by the the sparse LDA algorithm in the simultaneous feature selection and classification step of the final model.

Table 1. Original and sparse implementation of the PLS regression algorithm. Observe the sparse PLS algorithm does the iteration using the selected variables and updates all direction vectors on the subspace of the remaining variables. This procedure guarantees the orthogonality of the resulted space.

Original PLS algorithm	Sparse PLS algorithm
1 Y_0,=Y-mean(Y), S=x*Y_0, C=S*S	Execute the lines 01 to 04 from the original algorithm
2 for a = 1,. . . , A	
3 q = dominant eigenvector of C	1 zw=(abs(w)<=threshold)
4 w=S*q	2 w(zw)=0
5 t=X*w, t=t-mean(t)/sqrt (t*t)	3 C(zw)=0
6 w = w/norm	Continue from line 3 of the original algorithm
7 q=Y_0*t, p=X'*t, u=Y_0*q, v=p	
8 if a > 1	
9 v=v-V*(V *p), u=u-T*(T*u)	
10 v = v/SQRT(v *v), store v in V	
11 s = V-v*(v *s)	
12 Store w, t, p, q, and u into	
13 W, T, P, Q, and U, respectively	
14 end	

3.4 Classification and Evaluation

The performance of the methods was evaluated using a 10-fold, cross-validation approach. For classification, the framework used the Fisher LDA. For evaluation, we measured the area under the ROC curve (AUC). Since the data was unbalanced with respect to the number of knees for each class (healthy/OA), cost functions such as the classification accuracy were inappropriate.

4 Experiments and Results

To investigate the performance of the sparsity methods on the model accuracy, we performed five experiments in each dataset. The experiments applied different DR methods to the 534 original features generated in accordance with Section 3.2.

The first experiment applied the Fisher LDA considering the Moore-Penrose pseudo-inverse of the covariance matrix, since the number of samples was less than the number of the original features. The second experiment applied the SIMPLS algorithm of PLS regression. The DR step of the training phase defined the number of PLS latent factors to use in the final model. The next experiment evaluated the performance of the forward feature selection method presented in [8]. Likewise, the two last experiments evaluated the performance of the sparse PLS and the sparse LDA. Their outcomes were propagated to the final classification process. Table 2 compares all the evaluations.

Table 2. Evaluations of the different methods: Fisher LDA, PLS regression, PLS with forward feature selection (PLS-FFS), sparse PLS and sparse LDA. The third columns show the number of features used in each CV fold evaluation. The fourth columns show the number of features used across all CV folds evaluations.

Diagnosis				Prognosis			
Method	AUC	features per fold	features all folds	Method	AUC	features per fold	features all folds
LDA	0.86	534	534	LDA	0.63	534	534
PLS	0.88	534	534	PLS	0.67	534	534
PLS-FFS	0.89	116	212	PLS-FFS	0.69	43	108
SPLS	0.93	349	500	SPLS	0.70	44	56
SLDA	0.89	113	268	SLDA	0.59	136	306

5 Discussion and Conclusion

Comparing LDA and sparse LDA, the results reported in Table 2 indicated that sparse LDA improved the diagnosis evaluation and reduced considerably the final feature space. But contrary to expectations, there was no detectable improvement to the prognosis evaluations. One possible explanation can be overfitting, since during the training phase, the method had median AUC of 0.99 for both datasets.

However, by including sparsity in the PLS algorithm we could increase the model accuracy and identify the subset of features actually used by the texture analysis framework. In general, the sparse PLS performed better than all other evaluated methods. The accuracy improvement was more expressive in the diagnosis evaluation, where the AUC reached 0.93. Comparatively, a recent study analysing a linear combination of morphometric and structural cartilage markers in the same population scored AUC of 0.84 [14]. Although the studies analysed different anatomical structures, the results showed the sparse PLS captured the texture changes and had diagnostic ability superior to other biomarkers of OA.

In the prognosis, the sparse PLS reached an AUC of 0.70. Although the performance was only slightly better than the other methods, the number of features selected were only 9% of the available ones. Considering all cross validation folds, we can notice some overlap between the selected feature sets, indicating the stability of the model. The sparse PLS decreased the model complexity, which can potentially contribute to a better understanding of the anatomical characteristics of the data being analysed.

Future improvements include evaluating the framework with some robust sparse algorithms, where the algorithm identifies and treats outliers before distinguishing the relevant features. Furthermore, validation on another dataset is key for verifying the results.

In conclusion, we presented a investigation of sparsity methods for dimensionality reduction in texture analysis. The results illustrated that by including a sparsity approach, our framework limited the number of features used by the

final model and increased the performance ability of separating healthy and OA subjects.

Acknowledgments. We gratefully acknowledge the funding from the Danish Research Foundation (Den Danske Forskningsfond) supporting this work and the Center for Clinical and Basic Research for providing scans and radiographic readings.

References

1. Blumenkrantz, G., Lindsey, C.T., Dunn, T.C., Jin, H., Ries, M.D., Link, T.M., Steinbach, L.S., Majumdar, S.: A pilot, two-year longitudinal study of the interrelationship between trabecular bone and articular cartilage in the osteoarthritic knee. Osteoarthritis and Cartilage **12**(12) (2004) 997–1005
2. Schad L.R., Blml S, Z.I.: Mr tissue characterization of intracranial tumors by means of texture analysis. Magnetic Resonance Imaging **11**(6) (1993) 889–896
3. Kovalev, V.A., Kruggel, F., von Cramon, D.: Gender and age effects in structural brain asymmetry as measured by mri texture analysis. NeuroImage **19**(3) (2003) 895–905
4. Sørensen, L., B., S.S., de Bruijne, M.: Quantitative analysis of pulmonary emphysema using local binary patterns. IEEE Transactions on Medical Imaging **29**(2) (2010) 559–569
5. Filzmoser, P., Gschwandtner, M., Todorov, V.: Review of sparse methods in regression and classification with application to chemometrics. Journal of Chemometrics **26**(3-4) (2012) 42–51
6. Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. The Annals of Statistics **23**(1) (1995) pp. 73–102
7. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. Technometrics **53**(4) (2011) 406–413
8. Marques, J., Dam, E.: Texture analysis by a pls based method for combined feature extraction and selection. In: Proceedings of the Second international conference on Machine learning in medical imaging. MLMI'11, Berlin, Heidelberg, Springer-Verlag (2011) 109–116
9. Hyonho, C., Kele¸s, S.: Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(1) (2010) 3 – 25
10. Folkesson, J., Dam, E.B., Olsen, O.F., Pettersen, P.C., Christiansen, C.: Segmenting articular cartilage automatically using a voxel classification approach. IEEE Transactions on Medical Imaging **26** (2007) 106–115
11. Florack, L.M.J., Haar Romeny, B.M.t., Koenderink, J.J., Viergever, M.A.: The Gaussian scale-space paradigm and the multiscale local jet. International Journal of Computer Vision **18**(1) (1996) 61–75
12. de Jong, S.: Simpls: An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems **18**(3) (1993) 251–263
13. Shewchuk, J.R.: An introduction to the conjugate gradient method without the agonizing pain. Technical report, Pittsburgh, PA, USA (1994)
14. Dam, E.B., Folkesson, J., Pettersen, P.C., Christiansen, C.: Automatic morphometric cartilage quantification in the medial tibial plateau from mri for osteoarthritis grading. Osteoarthritis Cartilage **15**(7) (2007) 808–18